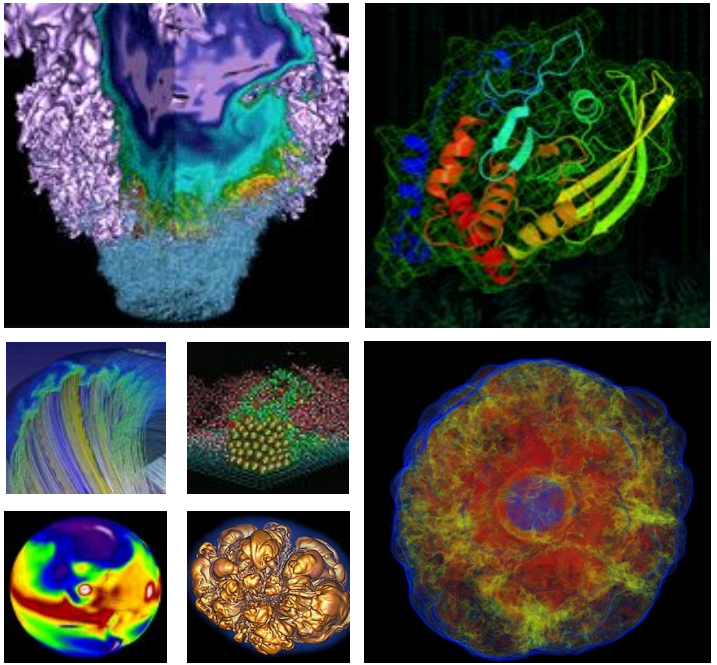
The NERSC Superfacility Project:

A Technical Overview



Cory Snavely

Lead, NERSC Infrastructure Services Group

NERSC GPUs for Science Day 1
2019-07-02
Lawrence Berkeley National Laboratory

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

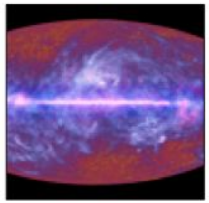
Agenda and Purpose

Describe the NERSC Superfacility Project

- Background from a Compute Facility Perspective
- Project Goals
- Science Engagements
- Technical Work Areas
- Example Outcome

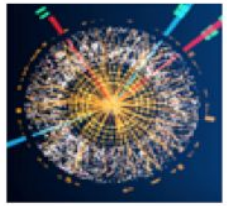
NERSC supports a large number of users and projects from DOE SC's experimental and observational facilities

Palomar Transient Factory
Supernova



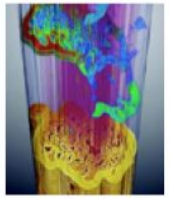
Planck Satellite
Cosmic Microwave Background
Radiation

Star
Particle Physics



Atlas
Large Hadron Collide

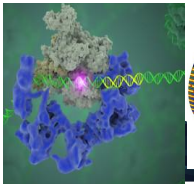
Dayabay
Neutrinos



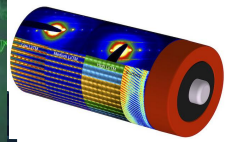
ALS
Light Source

LCLS
Light Source

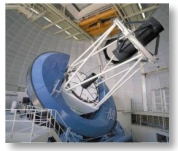
Joint Genome Institute
Bioinformatics



Cryo-EM

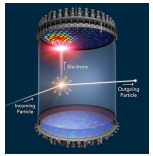


NCEM

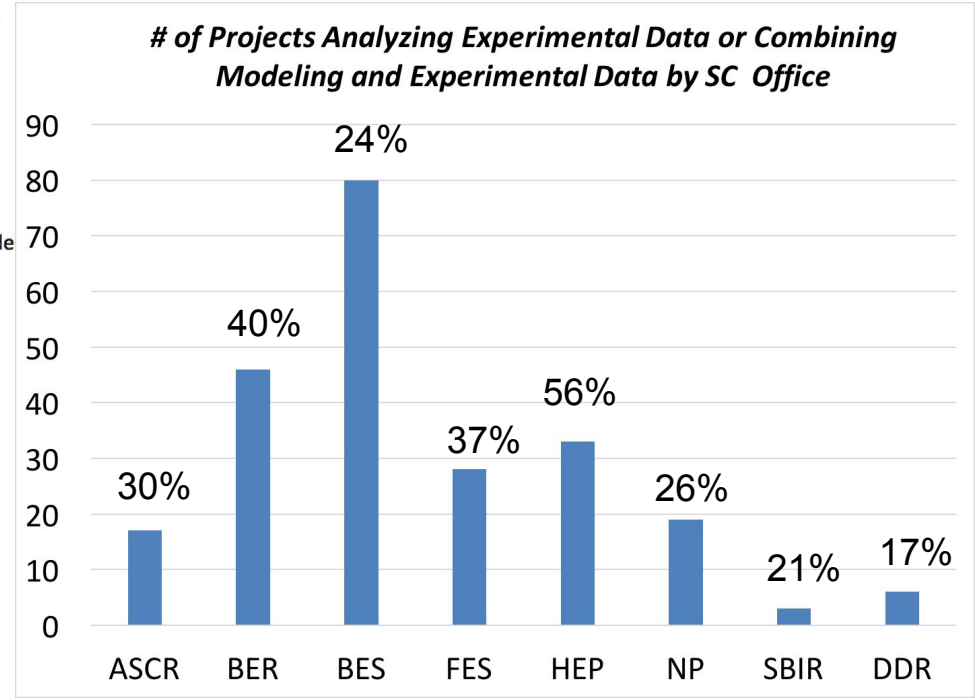


DESI

LSST-DESC



LZ



~35% (235) of ERCAP projects self identified as confirming the primary role of the project is to 1) analyze experimental data or; 2) create tools for experimental data analysis or; 3) combine experimental data with simulations and modeling

Themes from Requirements Reviews

- ***Scalable*** **methods for analysis and reduction of large datasets**
- **Large storage systems with *high performance* and *intuitive* interfaces**
- **Significant advances needed to *search, publish*, and *share* data**
- ***Seamless data movement*** **throughout workflow**
- **A *co-evolution of capabilities* between DOE experimental and ASCR facilities**

What We Mean by "Superfacility"

Su•per- *prefix.* Placed over (as abstraction); transcending

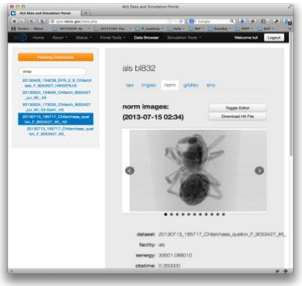
// Superstructure, superimpose, supersymmetry

An *ecosystem* of connected facilities, tools, and expertise to enable new modes of discovery.

Superman

Superfood

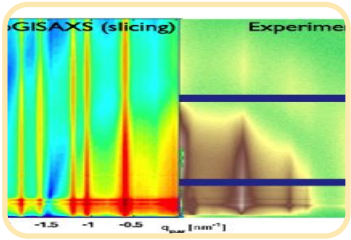
Superfacility: an ecosystem of connected facilities, software and expertise to enable new modes of discovery



Experimental Facilities

Real-time analysis and Data management

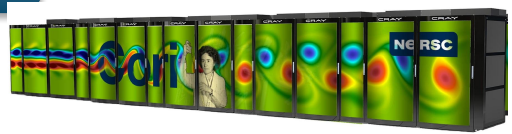
New Mathematical Analyses



Fast Implementations on latest computers

Computing Facilities

Network for Big Data Science

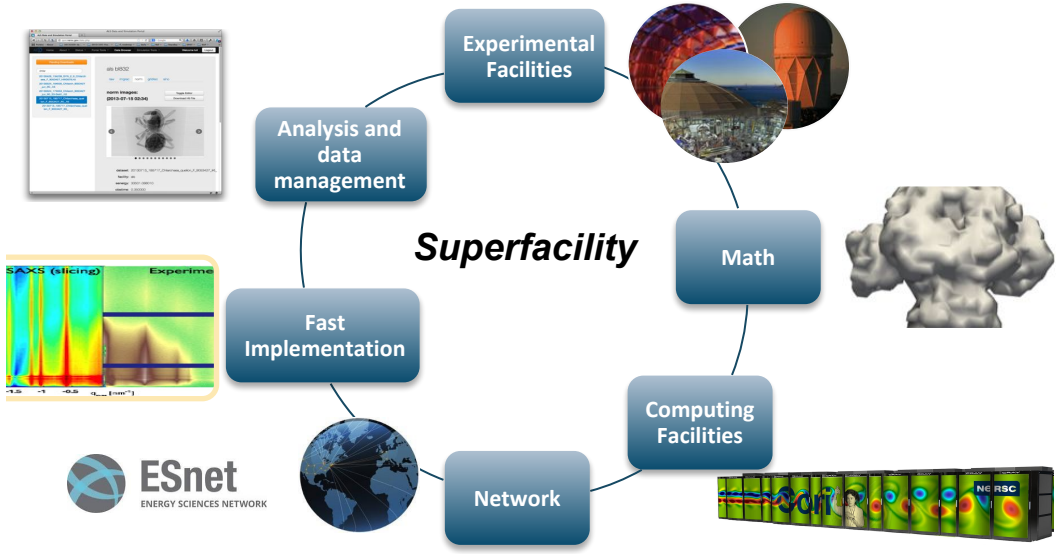


ESnet
ENERGY SCIENCES NETWORK

NERSC

What's the role of NERSC in a superfacility?

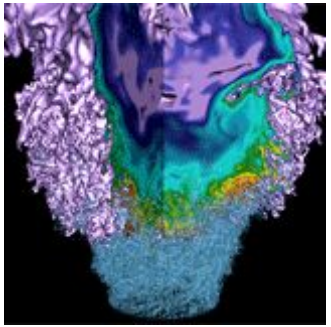
- Deploying large scale computing and storage resources
- Providing reusable building blocks for experimental scientists to build pipelines
- Providing scalable infrastructure to launch services
- Expertise on how to optimize pipelines

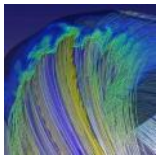


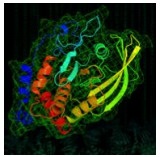
Superfacility Project Goals

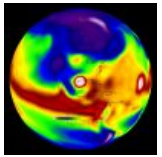
- First and foremost, demonstrate **science impact**.
- Provide **building blocks** that **lay the groundwork and help do heavy lifting** for complex workflows.
- Engage with a **breadth of projects** to satisfy different use cases. Aim: **generalized, reusable** outcomes.
- Emphasis on **distributed workflows**.
- **Coordinate existing efforts** under a common rubric for **maximum impact**.
- Involve NERSC peer divisions: **Computational Research, ESnet**.

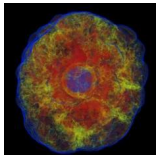
Science Engagements

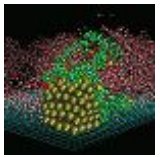












NERSC

Science Engagements

Just added!

Next-generation dark matter detection.

ALS

High-rate detectors use NERSC for real-time experimental feedback, data processing/management, and comparison to simulation

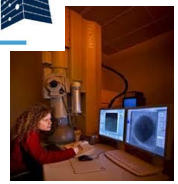
FICUS
Facilities Integrating Collaborations for User Science
JGI-EMSL Collaborative Science Call

Complex multi-stage workflow to analyse response of soil microbes to climate change

LSST DESC
Dark Energy Science Collaboration

Processing streaming alerts (from NCSA) for detection of supernova and transient gravitational lensing events

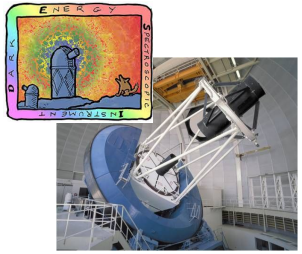
MOLECULAR FOUNDRY



4D STEM data streamed to NERSC, used to design ML algorithm for future deployment on FPGAs close to detector

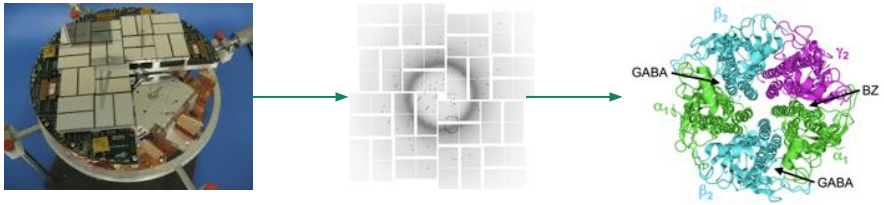
LCLS

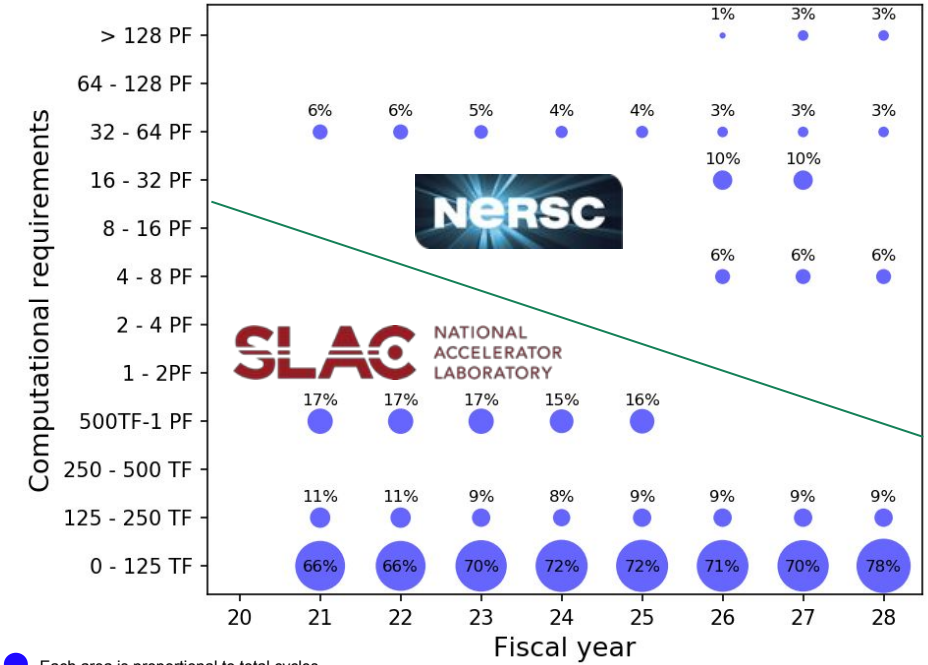
High-rate detectors use ESnet and NERSC for real-time experimental feedback and data processing



Nightly processing of galaxy spectra to inform next night's telescope targets

LCLS / ESnet / NERSC Collaboration

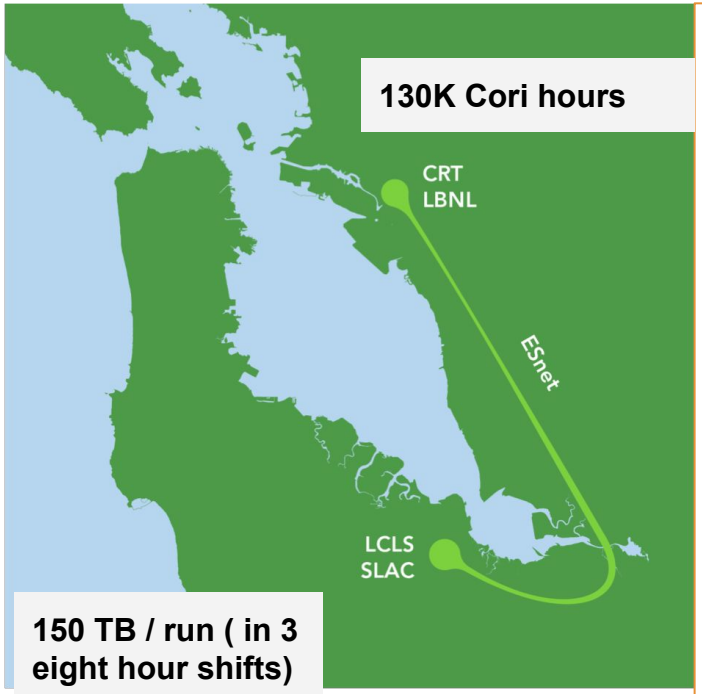




Science Summary:

- Streams of diffractive images reconstruct molecular structure and motion.
- Using HPC to speed data analysis allows on-shift understanding of collected data.
- 20% of LCLS-II (2021-2028) experiments will require NERSC (dots above line)

LCLS / ESnet / NERSC Collaboration



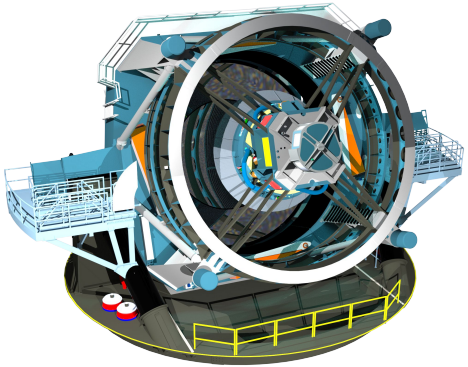
Needs from NERSC:

- Spin for data transfer automation
  - Reserve space and nodes for data
- Cori for Data Analysis
  - LCLS uses 130K hours per experiment
  - LCLS-II will 100x data rates
  - NERSC's ability to provide scheduled compute intensity is critical to this project
- GPUs for algorithm advancement
- WAN Bandwidth
  - In cooperation with ESnet provide scheduled bandwidth to compute nodes.
  - Orchestrate NERSC and ESnet resources (SENSE, SDN, scheduling)

LSST Dark Energy Science Collaboration

Science story

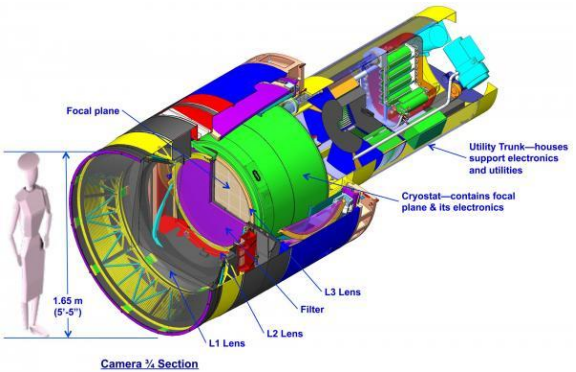
- Explain Dark Energy through multiple science probes: Galaxy catalogs, supernovae, lensing
- Survey covers the whole sky every few nights
- 3.2 Gpix camera built by DOE



Value proposition

Ability to co-locate and combine data w/compute:

- Simulations: Cosmology, instrument, detector
- Non-LSST Data: Other surveys for context
- Data analysis (HPC) and Sharing (Globus, Spin)

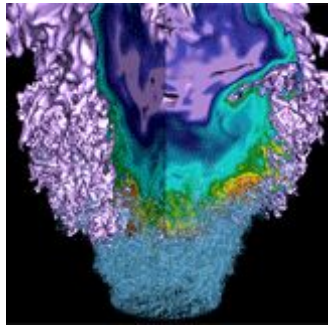


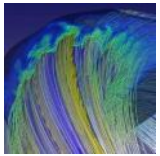
LSST Dark Energy Science Collaboration

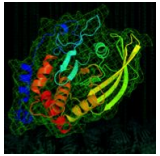
Needs from NERSC:

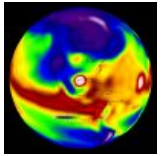
- Spin for Supernova broker
  - Nightly streaming data 27 MB/s,
- Cori for Simulations
  - 138M MPP hours in 2019 (increasing annually), 3 month scale turnaround
  - NESAP support
  - 1.2PB project storage purchased, additional 1PB in FY19
- Jupyter for analysis
  - Hundreds of scientists accessing notebooks

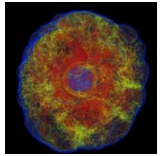
Technical Work Areas

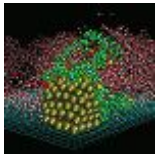












NERSC

Advanced Scheduling

Objective:

- Accommodate real-time jobs at significant scale

Tools/Techniques:

- Reservations made via API
- Checkpointing (requires support in codes)
- Preemption
- Queues / incentives
- SchedMD collaboration

Workflow Resiliency

Objectives:

- Greater general availability
- Machine-actionable alerts of service degradation

Tools/Techniques:

- Rolling upgrades
- Redundancy at minimally sufficient scale
- Build on the concept of “storage reservation”
- API-accessible center status

Scheduling Simulation

Objective:

- Quantify impact of scheduling changes

Tools/Techniques:

- Companion research project in LBNL Computational Research Division
- Baseline simulation with real test runs

Software-Defined Networking

Objectives:

- Provision capacity / connectivity on demand

Tools/Techniques:

- Companion efforts in ESnet / SENSE Project

Internal Data Movement

Objectives:

- Offload data movement between tiers (campaign, scratch, archive)

Tools/Techniques:

- Research and implement parallel data mover
- GPFS/HPSS Integration (GHI)

Data Dashboard

Objectives:

- Visualize storage usage
- Simplify management tasks ("archive this large directory")

Tools/Techniques:

- Data analysis using metadata from nightly scans

Spin (Edge Services)

Objectives:

- Simplify deployment of science gateways, workflow managers, databases, etc using containers

Tools/Techniques:

- Docker
- Compose (moving to Kubernetes)
- Rancher

API

Objectives:

- Comprehensive API endpoints for all of the above

Tools/Techniques:

- Kong frontend
- Backend workers
- Related: AWS, NERSC NEWT (legacy), TACC Agave

Federated Identity

Objectives:

- Share identities across facilities and services
- Simplify authentication, especially for distributed workflows

Tools/Techniques:

- OpenID Connect, OAuth 2.0, SAML
- JSON Web Tokens

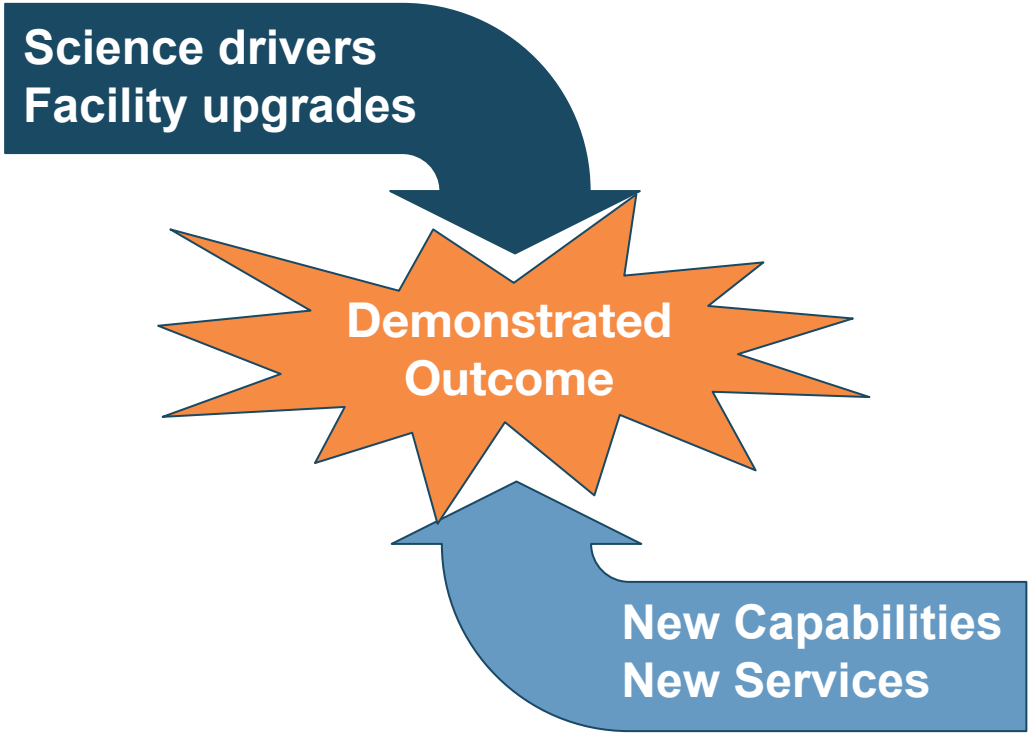
Intersections

	Advanced Scheduling	Resiliency	Scheduling Simulator	SDN	Data Movement	Data Dashboard	Spin	API into NERSC	Federated Identity
ALS	✓	✓	✓	✓	✓	✓	✓	✓	✓
DESI	✓	✓	✓		✓	✓	✓		
FICUS	✓	✓	✓		✓	✓	✓		✓
LCLS	✓		✓	✓				✓	✓
LSST-DESC	✓		✓	✓	✓		✓	✓	✓
NCEM	✓		✓	✓					

Bottom-up / Top-down Planning

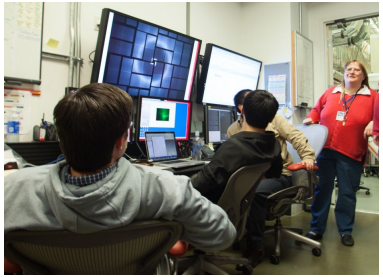
Novel workflows, facility and instrument upgrades present new challenges.

Engagement with a variety of projects ensures generalized solutions.



Example Science Goal

- **LCLS will test runs at NERSC this year to prepare for upgrading to LCLS-II in 2020.**



- **Currently requires many humans in the loop.**
  - NERSC staff to monitor and update compute reservation
  - ESnet and Network staff at SLAC/NERSC monitor data flows

- **Aim: simplify and automate the process.**

Example Science Goal

Use case: LCLS uses NERSC for a bursty experiment

- **Use API to make compute reservation and plumb SDN**
- **During run, use SDN to transfer data from experiment to compute nodes**
- **Launch and monitor compute jobs via a Jupyter notebook**
- **Results are automatically archived**

Smoother for user

- **Checkpointed code runs during reservation, using idle nodes**
- **Monitor the impact of the reservation on NERSC workload as a data point for the simulator**

Measured impact

Looking Forward

Project is Underway

- Kicked off in December 2018
- Three-year project
- Developing milestones at six-month intervals

Possible Future Directions

- Multiple compute facilities for resiliency?
- Data mirroring / migration?
- Workflow manager integration?

Thank You